HydraDAM2: Extending Fedora 4 and Hydra for Media Preservation

Jon W. Dunn, Will Cowan, and Juliet L. Hardesty, Indiana University
Karen Cariani, Rebecca Fraimow, and Sadie Roosa, WGBH Educational Foundation

March 30, 2018

The overarching goal of the HydraDAM2 project, funded by a grant from the National Endowment for the Humanities Preservation and Access Research and Development program, was to extend the existing HydraDAM digital asset management system, developed with prior NEH support, to be able to serve as a digital preservation repository for time-based media collections implementable at a wide range of institutions using multiple digital storage strategies. The new open source digital preservation repository system developed as part of the project by partners Indiana University (IU) and WGBH, known as Phydo, is based on the Fedora 4.x digital repository system and Samvera (formerly Hydra) repository application development framework and is intended to support storage and long-term preservation management of audio and video files and their accompanying metadata.

Project Activities and Accomplishments

The HydraDAM2 project sought to achieve five major goals:

1. Extend the HydraDAM digital asset management system to operate on the Fedora 4 repository system.
The original HydraDAM system developed by WGBH utilized the then-current Fedora 3 repository and was based on Sufia, a digital repository front-end application originally developed to support institutional research epositories. By the time the original HydraDAM project was completed in 2014, Fedora version 4 was ready for release and implemented major changes that had the potential to benefit the storage and preservation of large media files.

The project team's first tasks, planned during a kickoff meeting of IU and WGBH team members in Boston in February 2015, were to become familiar with the existing HydraDAM code and set up a Fedora 4 repository instance in order to determine whether it made sense to adapt the existing HydraDAM 1 code to work with Fedora 4 or to start over from a different code base. Given evolving changes in the Samvera codebase and fundamental differences between Fedora versions 3 and 4, the team elected to start development work over based on the Curation Concerns "gem," a Samvera-based front-end developed primarily in support of digital collections rather than institutional repositories.

IU and WGBH continued development of the Phydo system on Fedora 4 and Curation Concerns through 2016, with a major milestone being a release that was demonstrated by project directors Jon Dunn of IU and Karen Cariani of WGBH at the Open Repositories 2016 conference in Dublin, Ireland in June 2016. This release showed the ability to ingest content, provide metadata at the program and file level, search and browse on metadata, and execute file retrieval and fixity check operations against content stored in an external asynchronous storage environment such as a hierarchical storage management system using an automated tape library.

During the second half of 2016, development continued, focusing on: 1) integration of the specific HPSS (High Performance Storage System) hierarchical storage management system utilized by IU; 2) ingest, storage, and searching/browsing on preservation events stored as PREMIS[1]-compliant metadata; and 3) adapting the Phydo code base to use evolving versions of the underlying Hydra development stack. A second face-to-face meeting of the full IU/WGBH project team was held at Indiana University in Bloomington in September 2016, where the team focused on building out a feature roadmap for the remainder of the project[2] and working on issues related to implementation of PREMIS preservation events in the system.

As mentioned earlier, the original code base for the project was based on Curation Concerns and a unique data model. By 2016, the Samvera community was moving forward on two different fronts, one with the data model and the other with the code base. In early 2016, the Samvera community had chosen the Portland Community Data Model (PCDM) as the data model for Samvera development as the community moved forward. In 2016, it was obvious that Phydo would need to modify its data model to conform to PCDM or it would be outside the standards of the Samvera Community. This took considerable effort, but by early 2017 Phydo was using PCDM as its data model.

The second event involved the code base for Samvera. In the fall of 2016 the main discussion in the community was the merging of Curation Concerns and Sufia to become what was eventually known as Hyrax, the main code base for Samvera Application Development. Early in 2017, the team realized that it needed to upgrade Phydo's code to Hyrax and once again the team was in the position of devoting considerable effort to make this transition. By the summer of 2017 Phydo had been upgraded to Hyrax, and all development since then has been on this platform.

Telephone consultations with project advisory board members Hannah Frost of Stanford, Adam Wead of Penn State, and Andrew Woods of DuraSpace were conducted in September 2016, in order for team members to obtain advice on issues of content model design and integration of external storage.

[1] http://www.loc.gov/standards/premis/
[2] https://wiki.dlib.indiana.edu/display/HD2/Features+Roadmap+for+August+-+December+2016

2. Develop Fedora 4 content models for audio and video preservation objects, including descriptive, structural, and digital provenance metadata, based on current standards and best practices and utilizing new features in Fedora 4 for storage and indexing of RDF.
Through the evolution of its software development, Phydo team worked on moving the HydraDAM1 content model to align with the emerging Portland Common Data Model[3] (PCDM) from the Samvera and Fedora communities. Phydo is built as a Hyrax application using the Fedora 4 repository system. However, it was ultimately decided to continue to store audio and video files externally for performance reasons, and use Fedora to manage the metadata about the preservation files. Within Fedora, descriptive and structural metadata are stored about each of the preservation files, and preservation events are stored to record the digital provenance of the files.

Phydo's content model[4] utilizes Fedora 4 features for storage and indexing of metadata represented in the Resource Description Framework (RDF). The model incorporates Works, Filesets, and Files from PCDM. Descriptive and technical metadata properties used for discovery and access can then be managed as RDF attached to each of these types of objects. Structure is defined by relating Filesets that contain Files to Works. PREMIS events tracking preservation activities are associated with the Fileset so preservation actions can occur separately on different Files contained in a Work. The implementation of this model varied slightly between IU and WGBH in that WGBH does not make use of the Work as an object in its Phydo content model, only the Fileset and Files. IU includes pointers to the different types of audio and video files (preservation, production, access) and also includes multiple XML files for complete descriptive, technical, and provenance metadata. Each of these is a Fileset that IU groups together as a single Work in Fedora. WGBH does not include pointers to any digital files since its file storage uses physical LTO tapes stored offline, so the Fileset for WGBH contains the needed identifier information to physically locate the LTO tape, and the files managed in Fedora 4 are for technical and descriptive metadata per digital file. There is currently no need to further group by Work.

3. Implement support in HydraDAM for two different storage models, appropriate to different types of institutions: direct management of media files stored on spinning disk or on tape in a hierarchical storage management (HSM) system; and indirect management and tracking of media files stored offline on LTO tapes.
Phydo supports storage models that can accommodate the needs of both of these types of institutions. The release of Phydo that was demonstrated in June 2016 showed a first implementation of storage and retrieval of files on local or remotely-mounted server disk, as well as on asynchronous storage in an automated tape library, such as that at IU.

Integration of asynchronous storage is achieved via implementation of an *external file storage proxy* that allows files to be referenced in Fedora 4 that are stored in an external asynchronous

[3] https://github.com/duraspace/pcdm/wiki
[4] See https://wiki.dlib.indiana.edu/display/HD2/PHYDO+Data+Model for more detail.

storage system such as an automated tape library or high-latency cloud-based storage service. In IU's case, the storage proxy has been developed to work with the HPSS system used by IU to manage its Scholarly Data Archive (SDA) digital preservation and research data within an automated tape library environment mirrored between IU's Bloomington and Indianapolis campus data centers. The external storage proxy makes use of Fedora's API Extension Architecture (API-X) for extension of native Fedora features.

As part of the project, Indiana University and WGBH hosted a three-day "hackathon" in Bloomington, Indiana on September 10-12, 2017, with invited developers from other institutions within the Samvera and Fedora communities with needs for integration of asynchronous storage, including Duraspace, Stanford University, Art Institute of Chicago, University of Pennsylvania, Notre Dame University and Northwestern University, along with Phydo project team members from IU and WGBH. Hackathon participants worked collaboratively to define further requirements for the external file storage proxy and to work on code development. The results of the hackathon are documented on the project wiki.[5]

The external storage proxy architecture uses the Apache Camel message routing system in Fedora 4 to provide a route between the tape storage, Fedora, and a Samvera application such as Phydo. Objects are stored in Fedora 4 as non-RDF resources with a redirect URL so that the object determines which Camel route to use to download the file associated with the object. For this hackathon, the desire was to extend this architecture to include other external storage such as Amazon S3, Glacier, or other cloud services, and the hackathon participants chose to focus primarily on the Amazon S3 cloud storage service.

In the weeks leading up to the hackathon and over the course of the three days of the meeting, the Phydo team and hackathon participants focused on two major areas. One was defining and developing an API between the Hyrax user interface and the external storage proxy, and the second area was defining and implementing the exact structure of the external storage proxy. The design for the user interface API is complete and integrated with Phydo. The design and implementation in Camel of the File Storage Proxy is complete as well and integrated into Phydo. Users of Phydo may asynchronously download files from either tape or cloud storage.

For the offline storage model, WGBH implemented support for management of files stored on LTO tapes by developing a metadata tagging method within Phydo that allows WGBH archivists to tag large batches of ingested files on ingest with the LTO barcode identifier, which can then be mapped to shelf location through WGBH's MARS/PIM system for managing physical archival objects.

Additionally, WGBH developed methods to run a batch MD5 checksums on every file stored on a single LTO tape, then ingest those MD5s into Phydo for comparison against the MD5 checksums stored in Phydo metadata and automatically generate a PREMIS tag indicating

[5] https://wiki.dlib.indiana.edu/display/HD2/External+Storage+Proxy+Hackathon

whether the fixity check for that file failed or succeeded -- a key method for tracking file fixity in the files stored on the offline tapes.

4. Integrate HydraDAM into preservation workflows that feed access systems at IU (Avalon) and WGBH (OpenVault) and conduct testing of large files and high-throughput workflows.

For WGBH, Phydo was developed as an integrated part of existing preservation workflows for preserving materials generated by WGBH production units. These materials can include files as large as 700 GB for raw video footage in ultra-high definition. WGBH's existing preservation workflow utilizes Harvard's File Information Tool Set (FITS) to generate technical metadata about video files as a standard XML output, then stores the FITS XML in a Filemaker database, along with offline location information for the master files.

Over the course of the project, the WGBH team tested Phydo's ability to manage FITS XML describing the full range of media files created by WGBH's production units and extract key metadata values into the Fedora 4 repository, including file size, file format, frame size, and original file path, as well as tagging with location metadata on the batch level, as described above. The extraction of these key values allows WGBH staff to accurately identify production-quality files for re-use by production units at WGBH; it also allows staff to query against the Fedora 4 repository to gain information about file size and formats across the collection and use this data to make key preservation management decisions around questions such as file format obsolescence and digital storage migration. Although this information has not yet been linked to Open Vault, the extraction of metadata around identifiers such as file name and file path will support the eventual links between WGBH's preservation and access systems.

A typical batch of materials ingested from a WGBH production can include up to tens of thousands of individual files. Therefore, WGBH staff also tested high-throughput workflows by ingesting 16,000 FITS XML files at a time into Phydo using the hyrax-ingest gem. Through this testing, we discovered that the restrictions built into individual operating systems on simultaneously processing large numbers of files can impose limits on high-throughput workflows. We resolved this issue by improving the way the hyrax-ingest gem processed large batches of files to operate more efficiently and with fewer files open at once.

At IU, content is being ingested into its Phydo instance as an extension of existing workflows for the Media Digitization and Preservation Initiative (MDPI).[6] After media content has been digitized, the digital content along with technical metadata about the digital object are stored in the SDA hierarchical storage management system. The technical metadata needed for the preservation system, as well as pointers to the file locations for all digital content and metadata file, are extracted from SDA and ingested into the preservation system, Phydo. Since the digital content is being stored in SDA, Phydo contains important technical metadata and pointers to the location in SDA of the digital content. It is these pointers that the external storage proxy,

[6] https://mdpi.iu.edu/

discussed earlier, uses to access to the files on SDA. IU has done extensive testing with the external storage proxy to ensure that this access works correctly.

With several hundred thousand audio and video files digitized to date as part of MDPI and already stored in SDA prior to implementation of Phydo, IU is also concerned about the throughput from SDA to Phydo for ingestion. The IU team has a great deal of experience with ingesting content into Fedora 4 and will continue to tune the ingestion process. To date, IU has tested ingestion against several hundred files that deal with a cross section of digital media content digitized through MDPI workflows. All files digitized through MDPI are also transcoded and ingested into a private instance of Avalon Media System for access by collection managers and eventual promotion to on-campus or public access through Media Collections Online, IU's public instance of Avalon.

5. Document and disseminate information about our implementation and experience to the library, archive, digital repository, and audiovisual preservation communities.
Over the course of the project, project progress and outcomes were presented Open Repositories, Samvera Connect (formerly Hydra Connect), Association of Moving Image Archivists (AMIA), and the International Association of Sound and Audiovisual Archives (IASA). A complete list of presentations available online may be found in the Grant Products section below.

The project maintained its internal documentation in an open wiki site and Scrum board so that members of the advisory board and other interested parties were able to follow our work. All source code developed is available for reuse and contribution through the project's GitHub repositories, also listed in the Grant Products section of this report.

Development Process

Through the three-year project, WGBH and Indiana University (IU) worked together on technical design and code development to build Phydo, using an Agile Scrum software development methodology. Generally speaking, development was organized into a series of one- or two-week sprints, with developers from both institutions working together on user stories. Upon occasion, the teams would split and develop code specifically needed for only one of WGBH or IU, but in general, we found that working together was more productive. A decision was also made to keep the core product as close to the same as possible for IU and WGBH, only diverging when needed for the different storage scenarios needed by each: offline storage on LTO tape (WGBH) and nearline storage in a hierarchical storage management system (IU).

Over the last year of the project (2017), WGBH and IU focused development sprints on finalizing an implementable version of Phydo. As we got closer to the end of the project, in order to spur concrete progress, the sprint meetings began with a demo of code or functionality that had just been created. Specific development cycles worked on ingest, export, etc.

The project PI and Co-PI, along with other team members, continued to be active in the Samvera community, keeping abreast of developments in the core technical stack, and advocating when necessary for conscientious community progress to allow projects like Phydo to keep up with core developments. The project work was shared at meetings and conferences, and advice and input was welcomed from the Samvera community. Having a continued vital Samvera community is key in the continued sustaining needs for this project as it is based on core Samvera structures such as Hyrax and Fedora 4.

Audiences

Phydo is intended to serve the needs of libraries and archives with significant AV media collections, both digitized and born-digital, and to allow those libraries and archives to more easily manage the ingestion and ongoing preservation of AV files and associated metadata. By releasing Phydo under an Apache open source license, making the code publicly available through GitHub, and using technologies common to the library and archive communities, including the Ruby programming language, Ruby on Rails and Samvera development frameworks, and Fedora digital repository system, it is intended that the system may be adopted, adapted, and reused by a variety of institutions. In addition to adopting the entire Phydo system, various pieces of functionality developed as part of the project may be adopted on their own, including the *hyrax-preservation* gem that provides support for logging of preservation events using PREMIS data elements, the *hyrax-ingest* plugin that supports large-scale batch ingestion of metadata and content, and the asynchronous storage proxy discussed earlier.

Both IU and WGBH are in the process of adopting Phydo to support local digital preservation management needs, and the work of the project has been publicized so that other institutions with AV archival collections and/or with other preservation needs that align with the development work of Phydo may benefit from the system's development.

Evaluation

While the project was largely successful in meeting its objectives, the project did encounter several challenges. Some of the greatest challenges were related to continued change in several of the underlying technologies used to build Phydo, particularly within the Samvera Community (formerly Hydra Project). During the course of the project, significant changes were occurring both in the underlying content model used by Samvera and in its architecture and technology stack. Phydo development initially started before the creation of the Portland Common Data Model (PCDM) in late 2015, and so work had to be done to rework and adapt the application to use PCDM. Similarly, work on Phydo began with Sufia, then moved to various versions of Curation Concerns, and then to Hyrax, as the Samvera Community rapidly evolved its codebase and architecture. Late in the project, the team made a decision to pin its work to the most recent version of Hyrax (then Hyrax 1.0) in order to focus on completing feature

development for a Phydo 1.0 release rather than on keeping up with underlying technology stack changes.

As a result of these changes and our desire to keep current with the underlying technology stack, the team spent significant amounts of time adapting and rewriting features to keep up with new iterations of underlying dependencies. This is always a challenge for software projects that rely on third-party libraries to provide some of their functionality, but was a particular challenge for Phydo and other Samvera-based projects during the timeframe of this project. This challenge has been discussed and recognized within the wider Samvera community, and changes in development and release practices are being enacted to enable Samvera and Hyrax to serve as a more stable base for local software development projects.

The team also faced some challenges in developing effective collaborative workflows. Team members as individuals have varying working styles and methods; we learned that in order to synthesize the team into a productive unit, project managers at each institution had to not only be aware of the needs of their own developers, but also of the requirements of the developers at the collaborative institution. The project also faced turnover in project management and in the Scrum product owner role at both institutions midway through the project, which lengthened the process of discovering the collaborative working methods that would yield the highest productivity. Furthermore, because the project was designed to rely on a lean development team over a long timeline, changes in developer capacity due to illness, parental leave, or other events that could not have been anticipated at the beginning of the project caused significant delays in the project timeline.

In future development projects, we would recommend planning for a larger development team to focus on the project over a shorter timeline. This would result the project having greater stability without depending on the presence of any individual team member, and would also lessen the chances of upgrades in the underlying software dependencies overtaking the project before development was completed.

Continuation of the Project

Although the project has developed a tool that fulfills the basic needs for a digital preservation system, a full workflow still needs to be developed and tested in order to fully implement Phydo within WGBH's digital preservation environment. Moreover, systems throughout WGBH Educational Foundation are currently in a period of transition as WGBH makes a Foundation-wide shift to a new object store-based production storage system and more secure methods of accessing digital media content.

After the close of the grant and the launch of Phydo version 1.0, the WGBH project team will continue the work by testing workflows for ingesting large batches of new WGBH material, migrating records of existing material into Phydo, and running fixity checks on material that has

been ingested into past WGBH systems over the past three years. Workflows will be initially developed by the project team, and then tested and refined by WGBH archivists who have not been involved with Phydo development. We expect some further development will take place during this period as WGBH archivists test Phydo's functionality against their real-world use cases. The WGBH team will document these workflows for using Phydo in a production environment, and make them available through the project Wiki.

After this work has been completed, the WGBH team will migrate all remaining WGBH archival records into Phydo and transition to using Phydo to track digital preservation going forward. As WGBH IT continues development of the new production storage system, developers on the Phydo project team will collaborate with WGBH IT staff to build connections between Phydo and other existing WGBH-wide systems, including MARS/PIM, OpenVault, and WGBH Object Store.

Indiana University plans to continue work on integration of Phydo with the Avalon Media System access repository, enabling collection managers in Avalon to gain access to both the preservation master file content and technical metadata preserved in Phydo and stored in SDA, and vice-versa. In addition, IU will develop workflows for ingest of born-digital content into Phydo, SDA, and Avalon, alongside the digitized content from MDPI.

IU and WGBH will continue to communicate and collaborate regarding this additional Phydo development. Beyond that, the WGBH/IU partnership started on the Phydo project is also continuing through a new grant from the Andrew W. Mellon Foundation to WGBH, which will be working with consulting firm AVP to develop an improved access system for the American Archive of Public Broadcasting, with Indiana University advising from their experience implementing Avalon. In the long term, WGBH plans to implement an open-source access system for discovery of WGBH internal content that can be easily connected to Phydo. Avalon is one of tools currently in consideration for this open-source access system, which would provide an opportunity to further develop the partnership between WGBH and IU.

Long Term Impact

This project has allowed both IU and WGBH to have a digital preservation system for AV that is sustainable and part of a larger open source community. Both institutions can now effectively manage their digital preservation files and the technical metadata needed. Working across two institutions increased the skills of each partner's developers and project managers in building preservation and access systems and gave staff additional skills in managing agile development projects and effectively managing source code as part of a larger development project.

In addition, WGBH was able to remain active in the Samvera community and continue to push for software solutions and systems to manage digital audio visual materials. WGBH's work with IU led to another grant from the Andrew W. Mellon Foundation for implementation of Avalon as a metadata management system for the American Archive of Public Broadcasting. WGBH

hopes to further utilize Avalon for internal WGBH access to its media collection and connect it to the preservation files in Phydo.

IU plans to leverage code and experience gained in developing Phydo as part of its work to develop and implement a more general-purpose digital repository for ingest and preservation of born-digital archival content in all formats, as part of university-wide efforts to strengthen archival management and records management capabilities.

For the larger humanities and archival community, there is now an open source AV digital preservation solution that can be implemented using either local storage or robotic HSM systems. While potentially not immediately usable out of the box due to unique local considerations, Phydo may be implemented by institutions using in-house or contracted developer resources to adapt it to local workflows and digital storage environments. Lessons learned on how to manage a joint open source project and shared through conference presentations can help other organizations, and hopefully encourage them, to undergo similar endeavors. This project demonstrates that collaborations can be difficult but ultimately worthwhile through sharing of knowledge and expertise.

Phydo is part of the Samvera community and core code will be updated by IU and WGBH as the Samvera stack is updated. The code is available through GitHub and an open source license. Both WGBH and Indiana University are happy to continue to share their experience and insight on the code and the project solution with the larger community. In addition, as noted earlier, specific developments as part of the project, are packaged such that they may be reused by other Samvera institutions and projects as part of their own larger digital preservation or digital repository systems.

Grant Products

The following products were produced during the course of the project:

- Software:
  - Phydo: https://github.com/IUBLibTech/phydo
  - Hyrax ingest plugin: https://github.com/IUBLibTech/hyrax-ingest
  - Hyrax preservation gem: https://github.com/IUBLibTech/hyrax-preservation
- Web sites:
  - Phydo wiki: https://wiki.dlib.indiana.edu/display/HD2/
- Documentation:
  - PHYDO PCDM Model: https://wiki.dlib.indiana.edu/display/HD2/PHYDO+Data+Model
  - PHYDO External Storage Proxy (asynchronous storage): https://wiki.dlib.indiana.edu/display/HD2/External+Storage+Proxy+Hackathon
- Presentations, available at https://wiki.dlib.indiana.edu/display/HD2/Publications+and+Presentations:

- Karen Cariani and Jon W. Dunn. “HydraDAM2: Repository Challenges and Solutions for Large Media Files.” 11th International Conference on Open Repositories, Dublin, Ireland, June 2016.
- Heidi Dowding. “Process as Product: Modularizing Digital Preservation to Serve Diverse Needs.” IFLA WLIC 2016, Columbus, Ohio, August 2016.
- Heidi Dowding and Michael Muraszko. “HydraDAM2: Building Out Preservation at Scale in Hydra.” Poster presentation, iPRES 2016, Bern, Switzerland, October 2016.
- Heidi Dowding, “HydraDAM2 for Audiovisual Preservation.” Poster presentation, Hydra Connect 2016, Boston, October 2016.
- Dunn, Jon W. & Cariani, Karen. “Applying Repository Systems to Audiovisual Preservation.” Open Repositories 2017, Brisbane Australia, June 30, 2017.